LLM-CXR: Instruction-Finetuned LLM for CXR Image Understanding and Generation

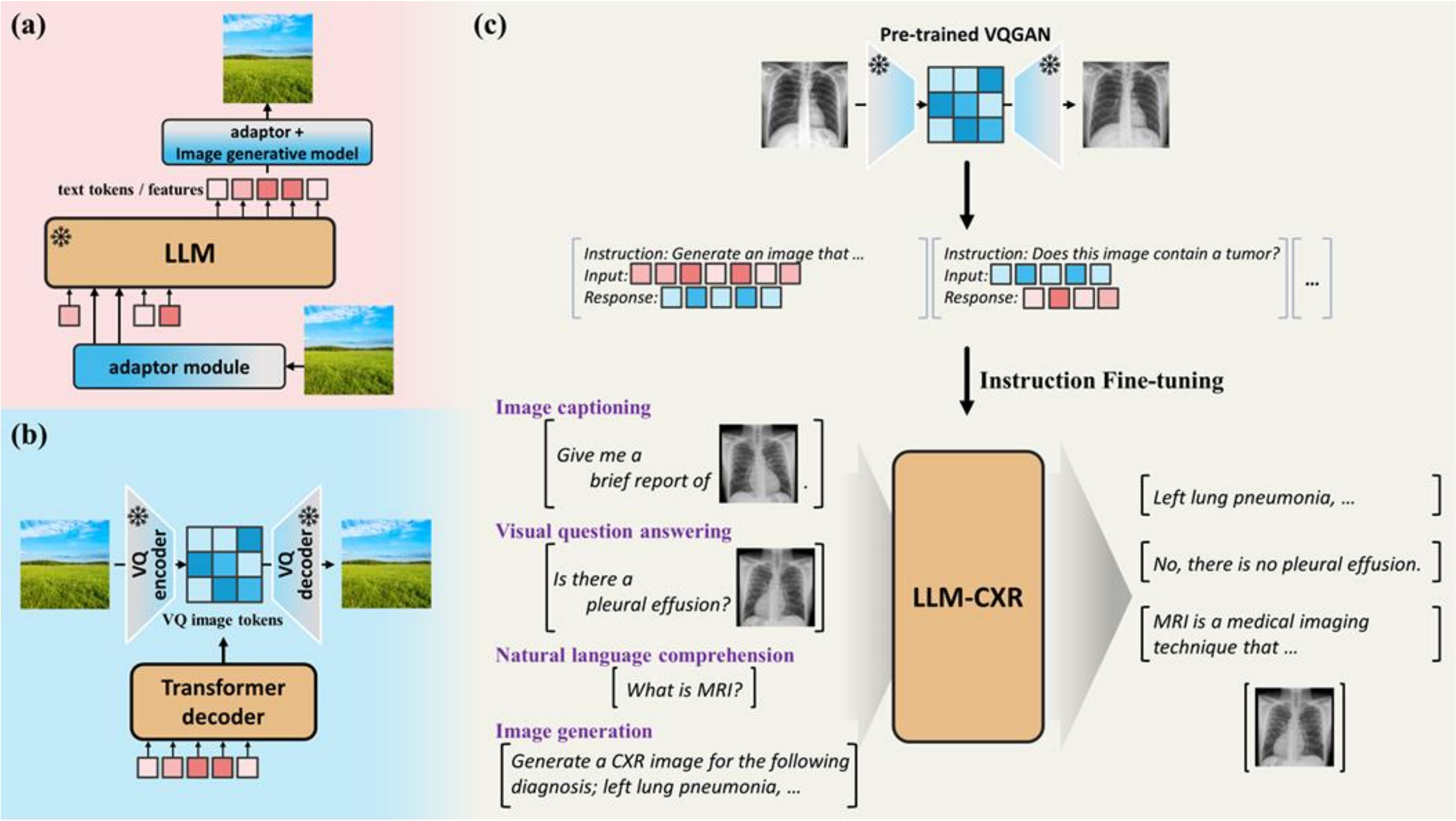
Suhyeon Lee*, Won Jun Kim*, Jinho Chang, Jong Chul Ye

Presenter: Won Jun Kim

Bio Imaging, Signal Processing & Learning Lab

Korea Advanced Institute of Science and Technology

Goal: LLM capable of CXR understanding and generation



Background

Multimodal Large Language Models

Example of GPT-4 visual input:

User — What is funny about this image? Describe it panel by panel.

Source: `https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/`

GPT-4 — The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

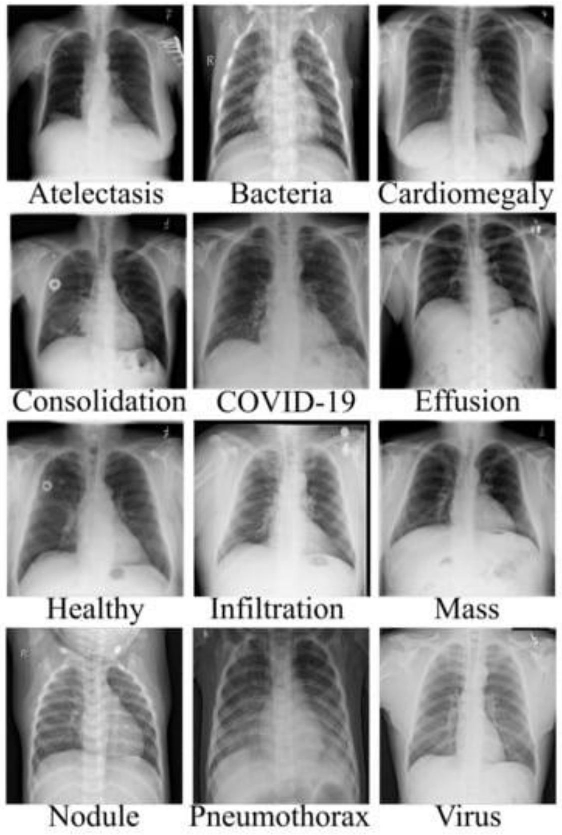
Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

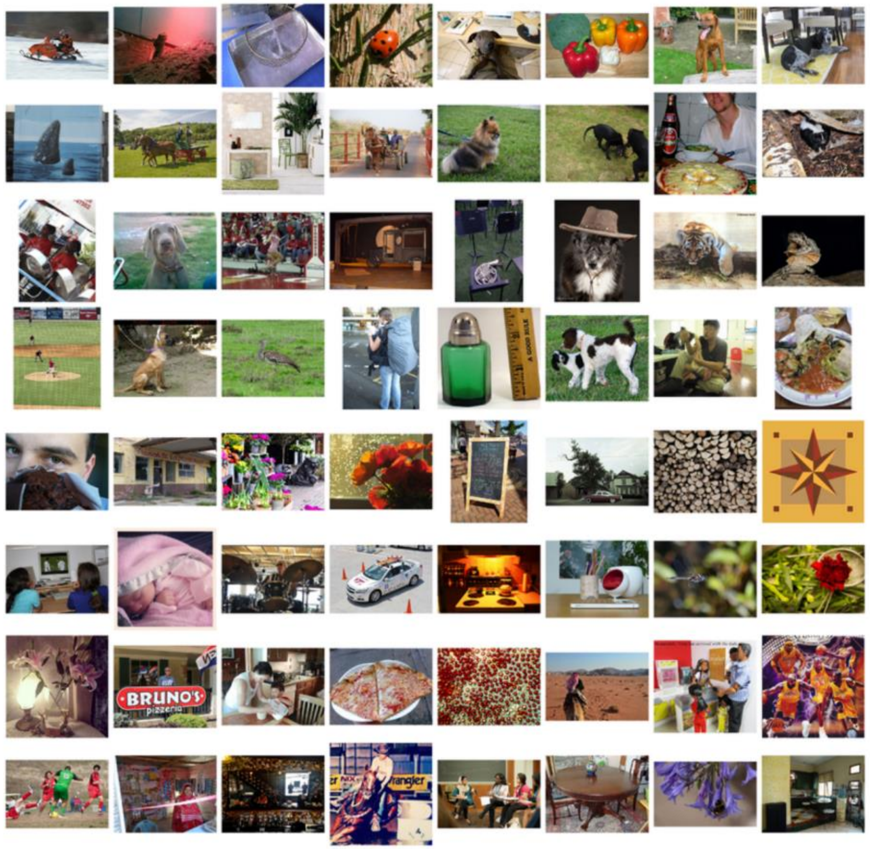
Table 3. Example prompt demonstrating GPT-4's visual input capability. The prompt consists of a question about an image with multiple panels which GPT-4 is able to answer.

[1] OpenAI, GPT-4 Technical Report.

Medical Vision-Language Alignment



[1]



[2]

[1] Monday et al., "COVID-19 Diagnosis from Chest X-ray Images Using a Robust Multi-Resolution Analysis Siamese Neural Network with Super-Resolution Convolutional Neural Network", *Diagnostics* 2022.
[2] ImageNet Large Scale Visual Recognition Challenge, 2015.

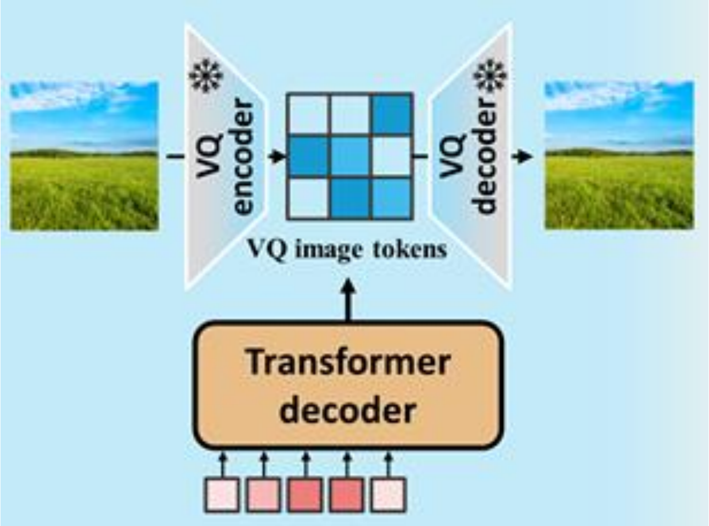
Previous work: Existing Approaches to CXR-understanding LLMs

- **RadFM[1]**
  - Images encoded through ***perceiver*** module (≈Flamingo[2])
- **ELIXR[3] / XrayGPT[4]**
  - Images encoded through ***Q-former*** module (≈BLIP-2[5])

[1] Wu et al., "Towards generalist Foundation Model for Radiology by Leveraging Web-scale 2D&3D Data", 2023.
[2] Alayrac et al., "Flamingo: a visual language model for few-shot learning", *NeurIPS* 2022.
[3] Xu et al, "ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiolo gy vision encoders", 2023.
[4] Thawakar et al, "XrayGPT: Chest radiographs summarization using medical vision-language models", 2023.
[5] Li et al, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models", 2023.

Previous work: Multimodal Transformers (Non-LLM)

- **Images** to token **embedding space** using a frozen **VQ-GAN** (VQ-VAE).
- Sequences of **text** and **image tokens** can be generated using an **autoregressive transformer** decoder



However, these models need to be trained from scratch with both image and text modalities from the beginning.

Our Contribution

- Proposal of a method for **fine-tuning** a **pre-trained LLM** using **images tokenized with a VQ-GAN** to achieve a well-aligned **bidirectional multimodal LLM**.

    - We leverage the **instruction-following capabilities** of a **pre-trained LLM**.
    - Give it **diverse instructions** for **CXR image understanding and generation** so that the instruction-tuning process achieves vision-language alignment.
    - We show that with our approach, we can train a **bidirectional multimodal LLM** that has **better vision-language alignment** than previous approaches while using a smaller base LLM.

Methods

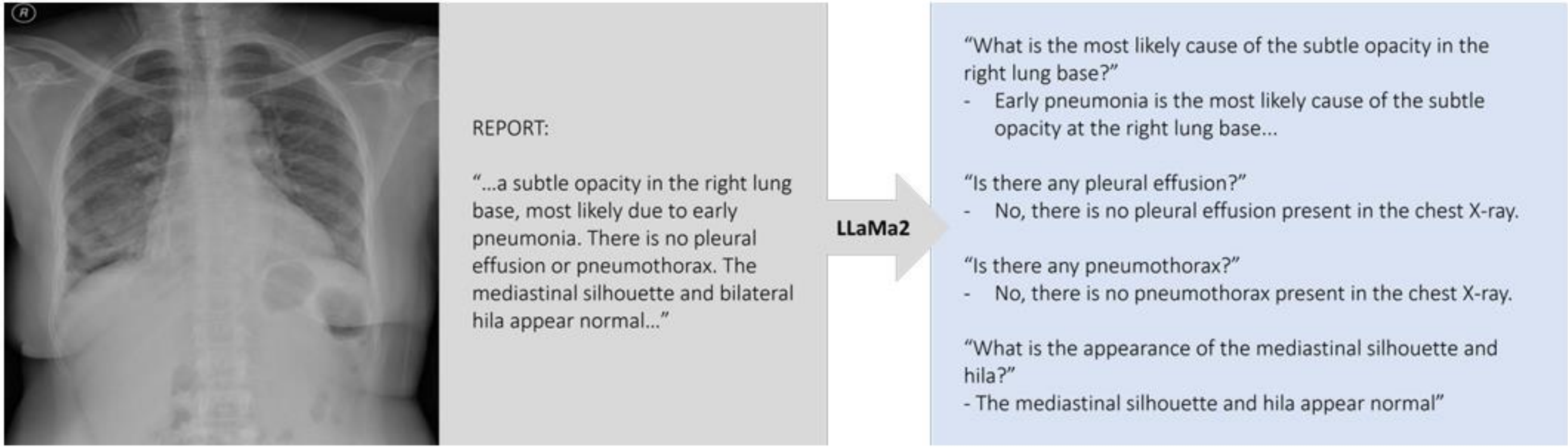
Tokenization of CXR Images

- **Images** are **tokenized** using a **VQ-GAN encoder** (decoded back to images using the corresponding decoder).

$$E(\cdot) : \mathbb{R}^{C \times H \times W} \rightarrow \{1, 2, ..., K_{\text{img}}\}^{d_z}$$
$$D(\cdot) : \{1, 2, ..., K_{\text{img}}\}^{d_z} \rightarrow \mathbb{R}^{C \times H \times W}$$

- Addition of **clinical information preserving loss** during VQ-GAN training.
  - 1024-dimensional feature ***L2 reconstruction loss extracted from a pretrained chest X-ray classifier***.

Synthetic Visual Question-Answering (VQA) Dataset



- **Text radiology reports** turned into **visual question-answer sets** using **LLaMa2-13b.**

Image-Text Bidirectional Instruction Fine-tuning

- Using the **instruction-finetuning scheme** used by the **Alpaca model family.**

Natural Language Instruction	*Report-to-CXR Generation*	*CXR-to-Report Generation*	*Visual Question-Answering*
###Instruction: Please summarize what LinkedIn does. **Input:** LinkedIn is a business and employment … **### Response:** LinkedIn is a social platform that businesses…	**### Instruction:** Generate a CXR image that corresponds to the following report. **Input:** Bilateral pulmonary opacities. **### Response:** <VQ032, VQ015, … VQ054, VQ032>	**### Instruction:** Generate a radiology report for the entered CXR image. **Input:** <VQ071, VQ056, …, VQ122, VQ002> **### Response:** No acute cardiopulmonary process.	**### Instruction:** What is the size of the bilateral pleural effusions in the image? **Input:** <VQ121, VQ070, …, VQ005, VQ428> **### Response:** Bilateral pleural effusions are moderate to large.

Training

- **Training objective**
  - **Loss** applied only to tokens **after** the ***Response*** **key**:

$$L_{instruct} = -\log p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{n_y} -\log p(y_i | y_{i-1}, y_{i-2}, \ldots, y_1, x_{n_x}, x_{n_x - 1}, \ldots, x_1)$$

- **Two-stage Training**
  - **First stage**: Unfiltered **high-volume** data.
  - **Second stage**: Filtered **high-quality** data.

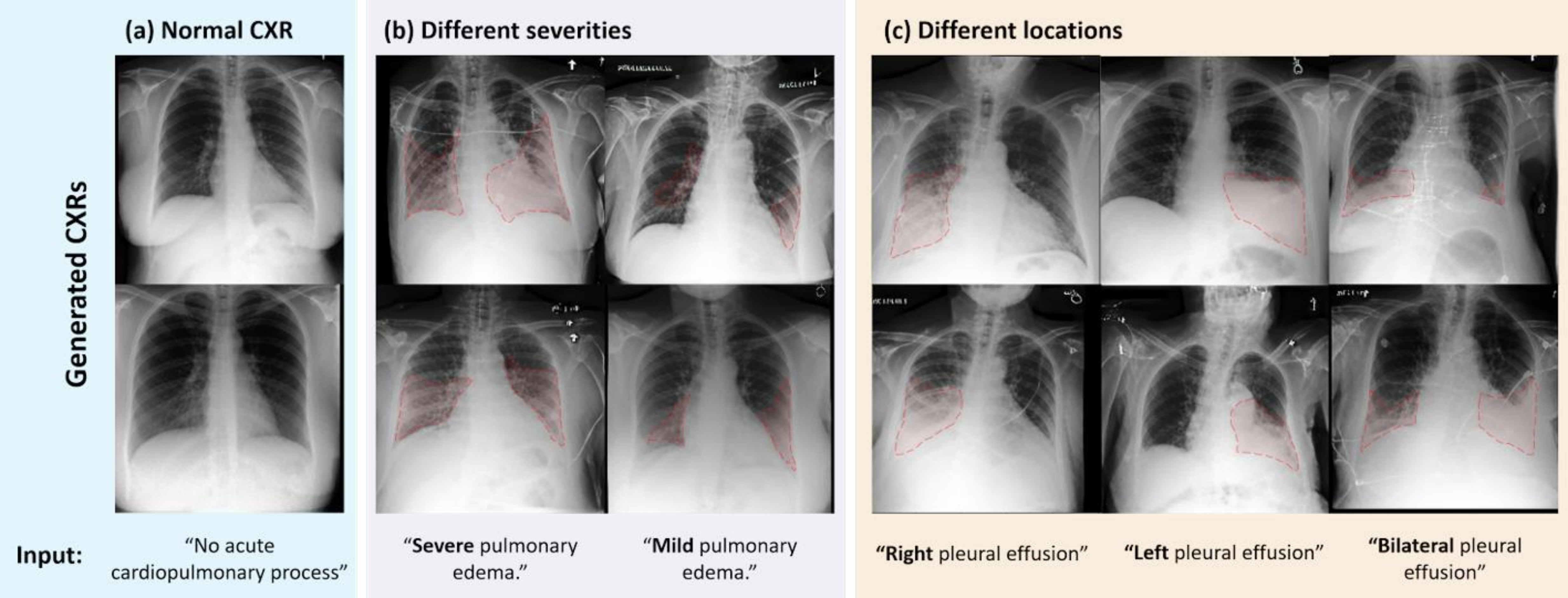
Results

CXR-to-Report (Qualitative)

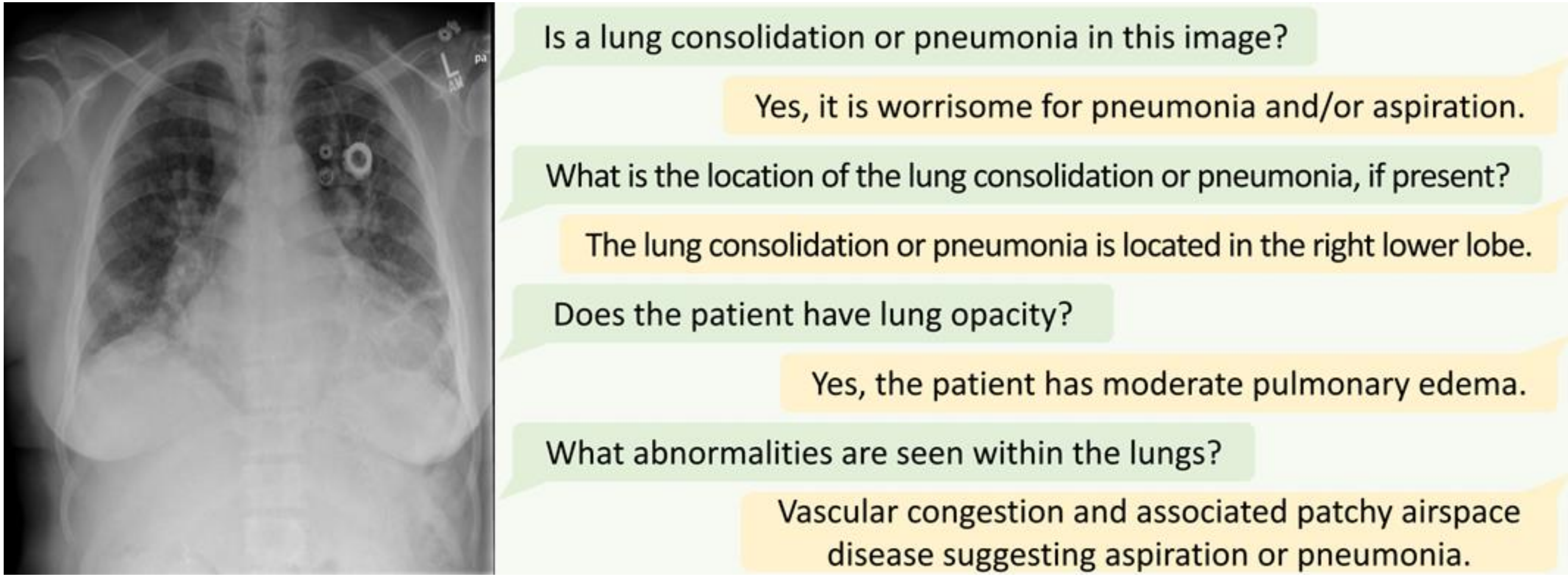
Input CXR

Ground-truth	"1. There is new moderate to large left pleural effusion. 2. Right pleural effusion is similar to prior.	"Left pleural effusion with overlying atelectasis. Left base opacity may be due to combination of pleural effusion and atelectasis... Mild pulmonary vascular congestion.	"Cardiac silhouette is enlarged but there is no vascular congestion. Opacification in retrocardiac region is consistent with volume loss in left lower lobe... difficult to exclude superimposed pneumonia...	"No evidence of acute cardiopulmonary disease."
Generated	"Bilateral pleural effusion."	"Mild pulmonary edema, pulmonary vascular congestion and small to moderate left pleural effusion."	"1. Mild cardiomegaly. 2. Left lower lobe consolidation could be atelectasis, pneumonia, or aspiration".	"No acute intrathoracic process"

Report-to-CXR (Qualitative)



VQA


L
Is a lung consolidation or pneumonia in this image?
Yes, it is worrisome for pneumonia and/or aspiration.
What is the location of the lung consolidation or pneumonia, if present?
The lung consolidation or pneumonia is located in the right lower lobe.
Does the patient have lung opacity?
Yes, the patient has moderate pulmonary edema.
What abnormalities are seen within the lungs?
Vascular congestion and associated patchy airspace disease suggesting aspiration or pneumonia.

CXR-to-Report (Quantitative)

Table 1: CXR-to-report generation AUROC and F1.[1]

AUROC ↑	Atel.	Cnsl.	Pmtx.	Edema	Eff.	Pna.	Cmgl.	Les.	Frac.	Opac.	ECm.	NoF.	P.O.	Dev.	Micro	Macro	Weighted
RadFM	0.587	0.498	0.503	0.633	0.657	0.504	0.611	0.516	0.498	0.514	0.502	0.666	0.499	0.597	0.638	0.556	0.596
UniXGen-512	0.570	0.533	0.519	0.615	0.682	0.526	0.645	0.501	0.498	0.555	0.510	0.676	0.498	0.740	0.668	0.576	0.628
IFCC	0.479	0.508	0.486	0.504	0.496	0.486	0.545	**0.518**	0.498	0.497	0.463	0.497	0.499	0.494	0.543	0.497	0.498
R2Gen	0.501	0.485	0.504	0.500	0.503	0.502	0.505	0.510	**0.500**	0.501	0.511	0.494	0.500	0.498	0.542	0.501	0.500
UniXGen-256	0.518	0.511	**0.530**	0.542	0.533	0.510	0.524	0.513	0.499	0.519	0.511	0.564	**0.527**	0.593	0.575	0.528	0.540
XrayGPT	0.551	0.506	0.511	0.590	0.595	**0.519**	0.570	0.511	0.499	**0.553**	**0.539**	0.592	0.490	**0.646**	0.617	0.548	0.577
LLM-CXR	**0.558**	**0.517**	0.496	**0.619**	**0.641**	0.509	**0.577**	0.506	0.494	0.537	0.505	**0.677**	0.498	0.640	**0.628**	**0.555**	**0.597**

F1 ↑	Atel.	Cnsl.	Pmtx.	Edema	Eff.	Pna.	Cmgl.	Les.	Frac.	Opac.	ECm.	NoF.	P.O.	Dev.	Micro	Macro	Weighted
RadFM	0.325	0.024	0.018	0.404	0.494	0.034	0.387	0.065	0.000	0.177	0.026	0.524	0.000	0.381	0.370	0.204	0.341
UniXGen-512	0.298	0.116	0.064	0.374	0.530	0.121	0.423	0.014	0.000	0.317	0.049	0.532	0.000	0.586	0.413	0.245	0.398
IFCC	0.159	**0.083**	0.020	0.203	0.312	0.006	0.270	**0.068**	0.000	0.323	0.042	0.200	0.000	0.292	0.220	0.141	0.225
R2Gen	0.168	0.018	0.020	0.073	0.129	0.034	0.263	0.043	0.000	0.240	0.051	0.289	0.000	0.254	0.201	0.113	0.183
UniXGen-256	0.146	0.072	**0.083**	0.226	0.215	0.072	0.176	0.055	0.000	0.282	0.047	0.411	**0.092**	0.367	0.262	0.160	0.243
XrayGPT	**0.279**	0.065	0.049	0.334	0.404	**0.110**	**0.347**	0.058	**0.016**	**0.352**	**0.076**	0.371	0.000	**0.470**	0.326	0.209	0.330
LLM-CXR	0.272	0.081	0.013	**0.382**	**0.464**	0.084	0.327	0.036	0.000	0.278	0.035	**0.535**	0.000	0.453	**0.360**	**0.211**	**0.350**

Report-to-CXR (Quantitative)

Table 4: CXR generation AUROC and F1.

AUROC ↑	**Atel.**	**Cnsl.**	**Pmtx.**	**Edema**	**Eff.**	**Pna.**	**Cmgl.**	**Les.**	**Frac.**	**Opac.**	**ECm.**	**Micro**	**Macro**	**Weighted**
RoentGen	0.7661	0.7535	0.6078	0.7084	**0.8169**	0.6054	0.7780	0.6283	0.6047	0.7162	0.7294	0.7061	0.7013	0.7055
UniXGen	0.7982	0.7509	0.6640	0.7876	0.7725	0.7065	0.7610	0.7200	0.7121	0.7867	0.7893	0.7435	0.7499	0.7518
LLM-CXR	**0.8054**	**0.8263**	**0.7540**	**0.8111**	0.8155	**0.7722**	**0.7846**	**0.7852**	**0.7596**	**0.8311**	**0.8335**	**0.7907**	**0.7980**	**0.7991**

F1 ↑	**Atel.**	**Cnsl.**	**Pmtx.**	**Edema**	**Eff.**	**Pna.**	**Cmgl.**	**Les.**	**Frac.**	**Opac.**	**ECm.**	**Micro**	**Macro**	**Weighted**
RoentGen	0.8113	0.7286	**0.7110**	0.2954	0.7619	0.2501	0.7639	0.2677	0.6580	0.7781	0.7066	0.6578	0.6121	0.6298
UniXGen	0.8648	0.6903	0.4981	0.7378	0.7008	0.7213	0.7598	0.5606	0.6424	0.7794	0.7958	0.7164	0.7046	0.7082
LLM-CXR	**0.8777**	**0.8283**	0.7024	**0.8061**	**0.8183**	**0.7529**	**0.8372**	**0.7678**	**0.7753**	**0.8342**	**0.8274**	**0.8065**	**0.8025**	**0.8054**

Thank You

Paper: https://arxiv.org/pdf/2305.11490

Code: https://github.com/hyn2028/llm-cxr